1.d

As is evident from $\Rightarrow w^* = \left(XX^T + \alpha I\right)^{-1} XY^T$
calculated above, this least sqaures problem becomes harder to compute as the matrices increase in size and become denser. Therefore, while linear regression offers a one step solution, because of being computationally expensive, iterative methods such as gradient descent need to be employed.

HW 2

HW 2.1: Human Activity Recognition Using Smartphones dataset

First dataset is cleaned by converting/combining Y labels to binary values as follows:

- -1 for walking (WALKING,WALKING UPSTAIRS,WALKING DOWNSTAIRS), and
- 1 for not (SITTING,STANDING,LAYING).

If x_i are column vectors,

$$\Delta f_i(w) = \frac{-y_i e^{-y_i w^T x_i}}{1+e^{-y_i w^T x_i}} x_i$$
$$\Delta f(w) = \frac{1}{N} \sum_{i \in [N]} \left(\frac{-y_i e^{-y_i w^T x_i}}{1+e^{-y_i w^T x_i}} x_i \right) + 2\lambda w$$

$$\Delta^2 f_i(w) = \frac{y_i^2 e^{-y_i w^T x_i}}{(1+e^{-y_i w^T x_i})^2} xx^T$$
$$\Delta^2 f(w) = \frac{1}{N} \sum_{i \in [N]} \left(\frac{y_i^2 e^{-y_i w^T x_i}}{(1+e^{-y_i w^T x_i})^2} xx^T \right) + 2\lambda \mathrm{I}$$

2.1 (a)

The quadratic regulariser term of $f(w)$ is not Lipschitz continuous, hence $f(w)$ is also not Lipschitz continuous

2.1 (b)

Note that the gradient of f_i is defined everywhere. Further, Δf_i is Lipschitz because $\Delta^2 f_i$ is upperbounded. Hence f_i is smooth.

Smoothness constant L is the supremum of the Hessian. Also, in ideal scenario, $w^T x_i = y_i$. Thus,
$$L_i = Max \left\{ \frac{y_i^2 e^{-y_i^2}}{(1+e^{-y_i^2})^2} x_i x_i^T \right\} = \frac{1}{2} ||x_i||^2$$

The function f is also smooth because the gradient of the regulariser is defined everywhere and its hessian is a constant matrix.

2.1 (c)

We have$\Delta^2 f_i(w) > 0,\ \forall x \neq \vec{0}$ (Positive Definite).
Let $x_1 = 0 \neq x_2, g(x) = xx^T$
$\Rightarrow g(px_1 + (1-p)x_2) = g((1-p)x_2) = (1-p)^2 x_2 x_2^T$
$< (1-p)g(x_2) = p.0 + (1-p)g(x_2) = pg(x_1) + (1-p)g(x_2)$
Thus $xx^T > 0$ for $x = \vec{0}$ as well.
⇒ $\Delta^2 f_i(w)$ is positive definite ⇒ Strong convex.

Since f is the sum of N strongly convex $f_i, i \in [N]$, f is also strongly convex.

Thus $\mu = 2\lambda + \text{Min.EigenValue of } \left(\frac{1}{N} \sum_{i \in [N]} \left(\frac{y_i^2 e^{-y_i w^T x_i}}{(1+e^{-y_i w^T x_i})^2} xx^T \right) \right)$, the second part of which can easily be found from the dataset

$\Rightarrow \boxed{\mu = 2\lambda + 5 \times 10^{-15}}$

Note: Eigenvalues depends on Y. If $Y \in \{0,1\}$ instead of $Y \in \{-1,1\}$, then $B = 2\lambda + 3 \times 10^{-15}$

Problem 2.2.

Since $EX^2 = Var[X] + [EX]^2$, we have

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[\|g(w_k;\xi_k)\|_2^2] \\
&= Var[g(w_k,\xi_k)] + E^2[g(w_k,\xi_k)] \\
&\overset{(2)}{\le} M + M_V\|\nabla f(w_k)\|_2^2 + E^2[g(w_k,\xi_k)] \\
&\overset{(1b)}{\le} M + (M_V + C_0^2)\|\nabla f(w_k)\|_2^2
\end{aligned}
$$

The proof ends by setting $M = \alpha$, $\beta = M_V + C_0^2$.

Problem 2.3.

Assume 1) smoothness

2) unbiased gradient

3) $Var[g(w_k,\xi_k)] \le M$.

$$W_{k+1} = W_k - \alpha_k g(w_k,\xi_k). \quad (2a)$$

Since f is L-smooth, we have

$$f(w_{k+1}) \le f(w_k) + \langle \nabla f(w_k), w_{k+1} - w_k\rangle + \frac{L}{2}\|w_{k+1} - w_k\|^2 \quad (2b)$$

Substituting (2a) into (2b), we have

$$f(w_{k+1}) \le f(w_k) + \langle \nabla f(w_k), -\alpha_k g(w_k,\xi_k)\rangle + \frac{L}{2}\alpha_k^2\|g(w_k,\xi_k)\|^2 \quad (2c)$$

Taking expectation on both sides of (2c) w.r.p. ξ_k, it gives

$$E f(w_{k+1}) \le E f(w_k) - \alpha_k\|\nabla f(w_k)\|^2 + \frac{L}{2}\alpha_k^2 E\|g(w_k,\xi_k)\|^2 \quad (2d)$$

Since $E\|g(w_k,\xi_k)\|^2 = Var[g(w_k,\xi_k)] + E^2[g(w_k,\xi_k)] = Var[g(w_k,\xi_k)] \le M$, it follows

$$E f(w_{k+1}) \le E f(w_k) - \alpha_k\|\nabla f(w_k)\|^2 + \frac{L}{2}\alpha_k^2 M. \quad (2e)$$

Rearranging (2e),

$$\alpha_k\|\nabla f(w_k)\|^2 \le f(w_k) - f(w_{k+1}) + \frac{L}{2}\alpha_k^2 M. \quad (2f)$$

Summing up (2f) over $k=1,\cdots K$, $\sum_k \alpha_k\|\nabla f(w_k)\|^2 \le f(w_0) + \frac{LM}{2}\sum_k \alpha_k^2 < \infty$.